2021

Managing your data & research objects for beginners

Thea P. Atwood
University of Massachusetts Amherst

Erin Jerome
University of Massachusetts Amherst

Managing your data & research objects for beginners

Thea Atwood
Erin Jerome

Intended learning outcomes

By the end of this session, you should be able to:

- Identify and describe specific actions you can take to enhance your data management practices.
- Compare your current data practices with good data practices and report on discrepancies.
- Identify resources to support practice change.

Data Management Resources @ UMass Amherst

Workshops - UMass Libraries > News & Events > Full events calendar

Web guide - Managing Your Data - https://guides.library.umass.edu/data

Consultations - email us! DWG@library.umass.edu

We can help with creating a data management plan, setting up file organization, trying an electronic lab notebook, finding data repositories, using the UMass Data Repository, and more.

Tips we'll cover:

1. Backups
2. File names & organization
3. Take good notes
4. Security
5. Future proof

Backups

Mike Bilderֵ, CEM @Bilder_CEM · Mar 15, 2019

Replying to @jimmyc42

When I coded for my **thesis**, I basically kept a Prof. Henry Jones Sr.-style Grail Diary that meticulously tracked every little thing I did in case I came back to it someday. In my recent move, I **lost** th all of my data and coding, plus the diary 🙃

1

Kevin @Kkbude · Mar 7, 2019

How to cry:
Put your **thesis** file at **flash drive** then **lost** it

1 1

LILITH @SAINTLILITHHH · Feb 7, 2019

i **lost** my **flash drive** with my whole **thesis** on it

sydney @sydneyslewis · Oct 30, 2018

hahahahaha i **lost** my **flash drive**, which had ev my master's program, including all the work for **thesis**!!!!! hahahahaha life is good y'all!!!!

1 6

Reward for stolen Royal Oak wreck data laptop

28 January 2019

HMS Royal Oak sank in 1939 after being torpedoed by a German U-boat

Divers and academics have offered a £1,500 reward for the safe return of laptop containing precious survey data from a sunken battleship.

Memorial University student's laptop containing master's thesis stolen

CBC News · Posted: Mar 08, 2016 5:50 PM NT | Last Updated: March 8, 2016

MUN graduate student Mark Colbourne had his laptop — and his master's thesis — stolen from his office. (Anthony Germain/CBC)

It's not a case of the dog eating your homework, but one student at Memorial University has lost his master's thesis — along with the laptop it was stored on.

Good practice: here, near, and far away

Here

A local or working copy.
E.g., on your workstation or in a shared workspace

Near

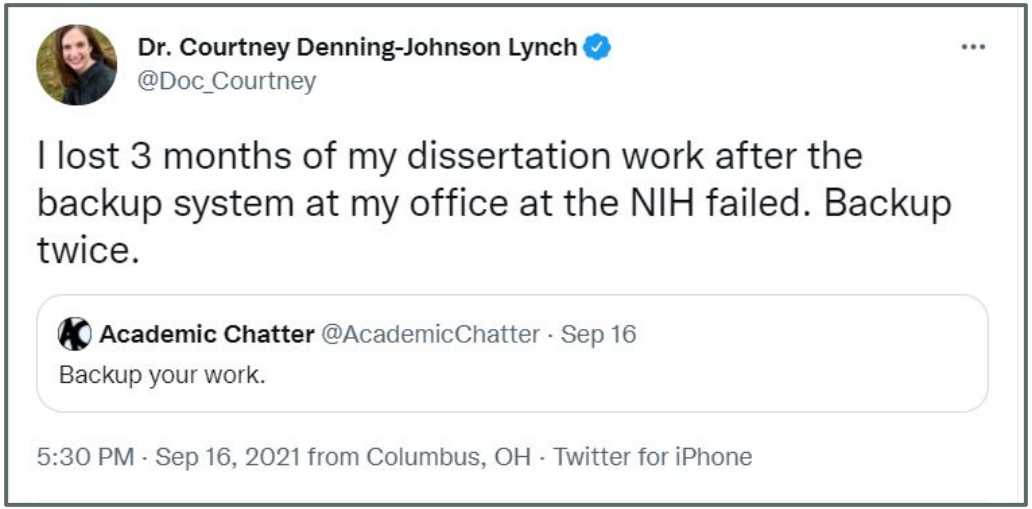
Another local or external copy
E.g., external hard drive (CDs and DVDs are not built to last)

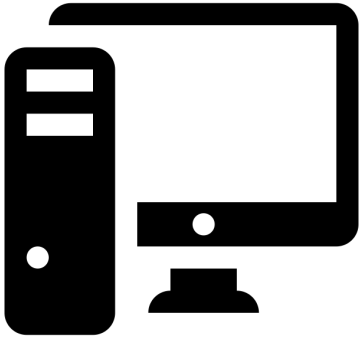
Far

A remote copy
E.g., in the cloud, like OneDrive.

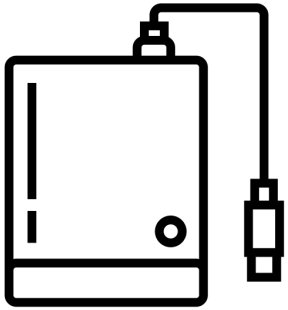
Test your file recovery system at setup and on a regular schedule.

Good practice: Here, near, and far away

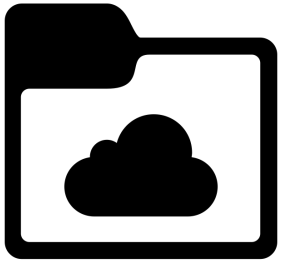




Created by icon 54
from Noun Project



Created by Creative Mania
from Noun Project



Created by Fabián Alexis
from Noun Project

Here

Your lab workstation

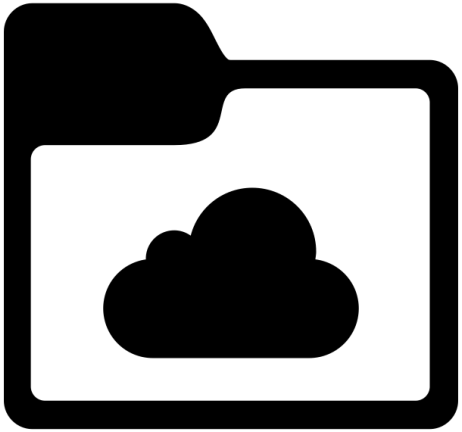
Near

An external hard drive

Far

OneDrive (‘the cloud’)

“The cloud is just someone else’s computer”



Created by Fabián Alexis
from Noun Project

Automated backup tools @ UMass Amherst

Time machine is the best and easiest way to backup Macs.

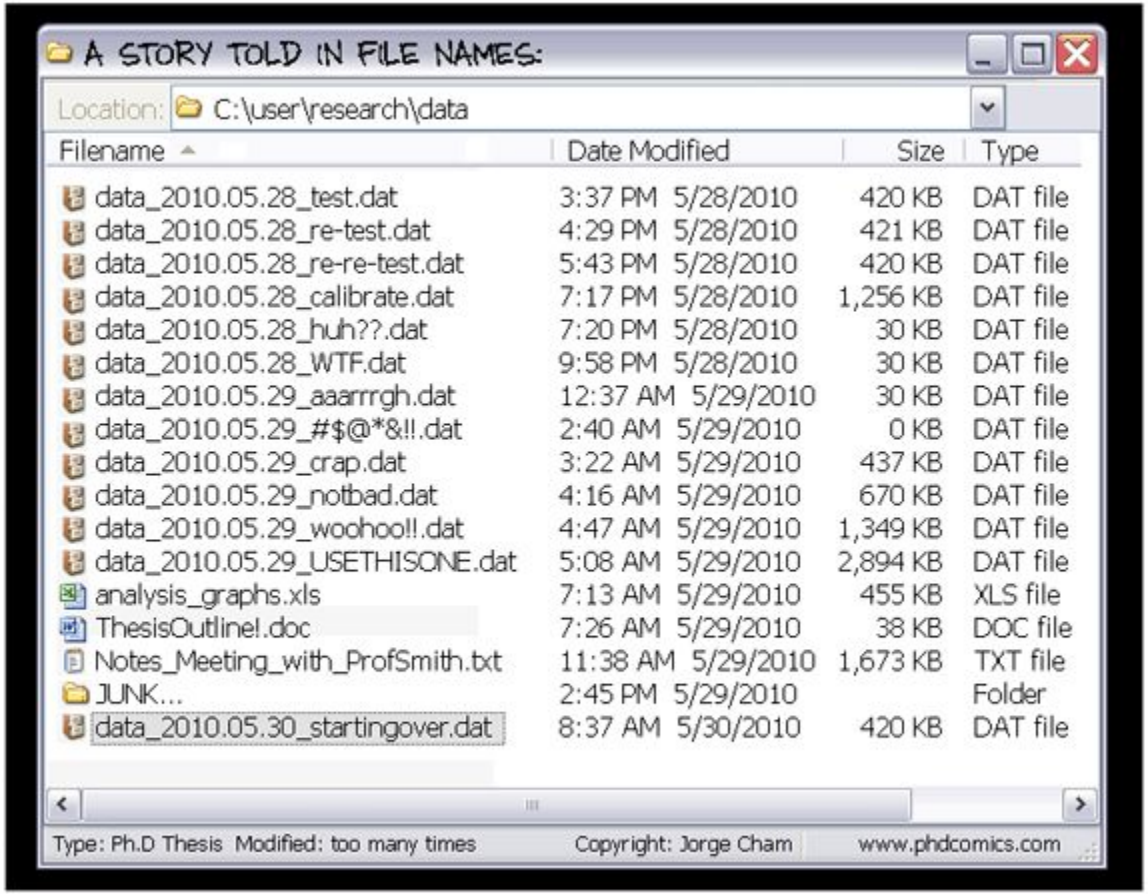
For Windows and Macs, UMass offers OneDrive which can be setup to backup to the cloud. More information is available here:

https://www.umass.edu/it/onedrive/video-guides

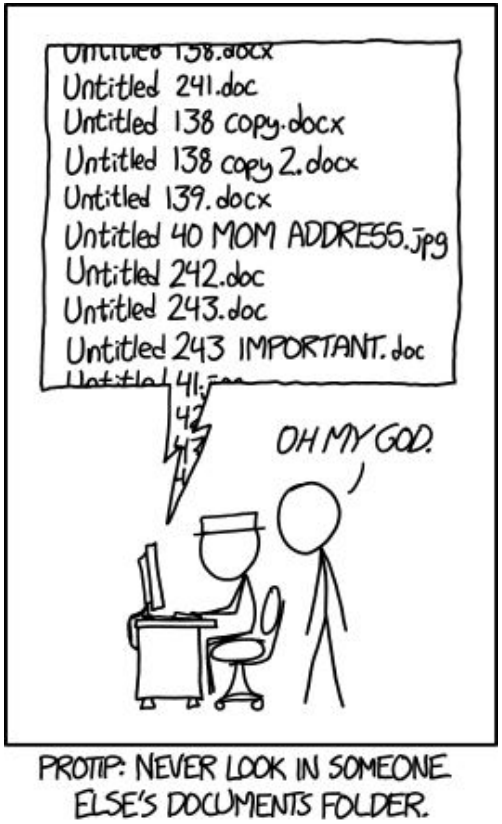
Data storage resources

- More info on OneDrive: https://www.umass.edu/it/services/onedrive
- More info on Google Apps: https://www.umass.edu/it/googleapps
- General info on where to store and share data: https://www.umass.edu/it/security/service-categorizations
- Info for when you leave UMass (~6 mo before onedrive, google drive, etc. data are deactivated & removed; ~ 1 year before google mail): https://www.umass.edu/it/accounts/information-students-leaving-umass-amherst

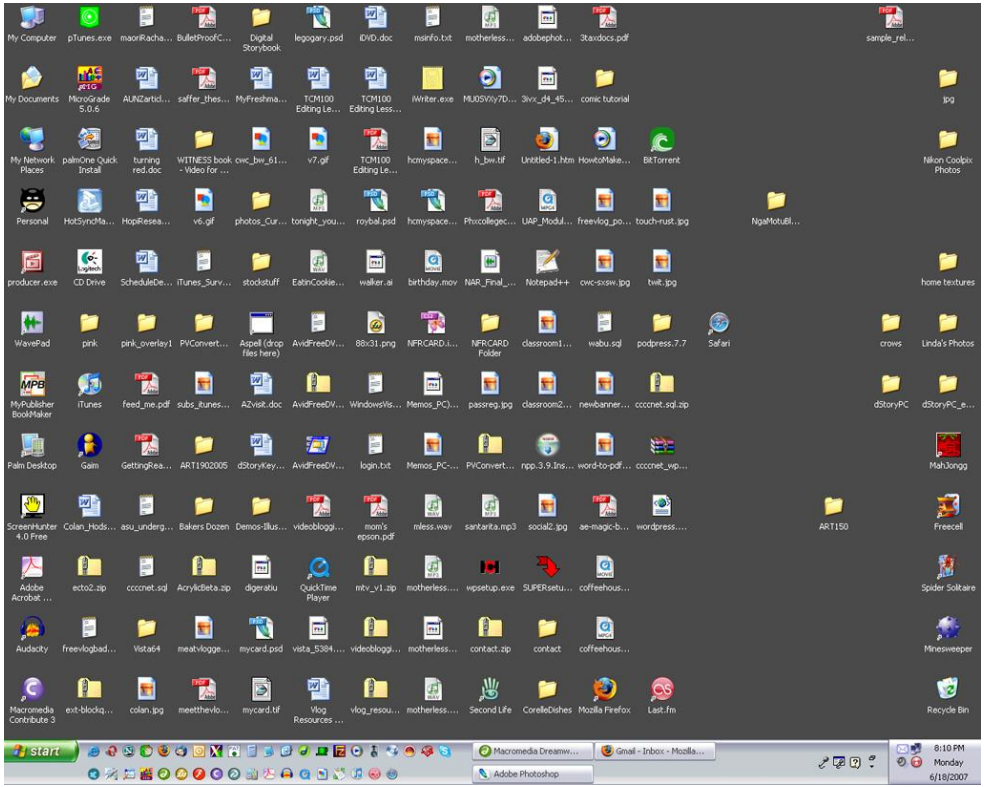
File names & organization



Cham, Jorge. (2010). “A story told in file names.” phdcomics.com. Available at http://phdcomics.com/comics/archive.php?comicid=1323



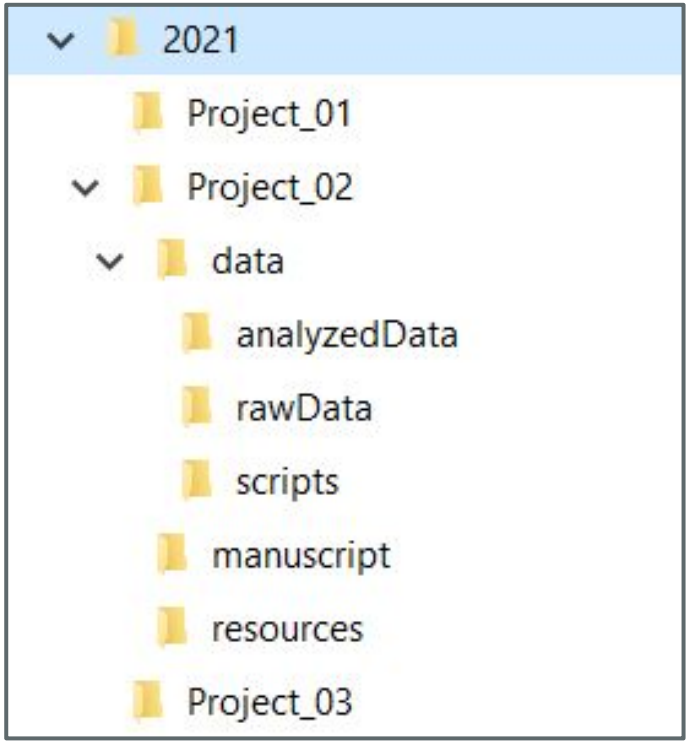
Munroe, Randall. (n.d.) "Documents." xkcd. Available at https://xkcd.com/1459/



Flickr user hummingcrows, (2007). "Before." Flickr.com. Available at https://flic.kr/p/SZdRG

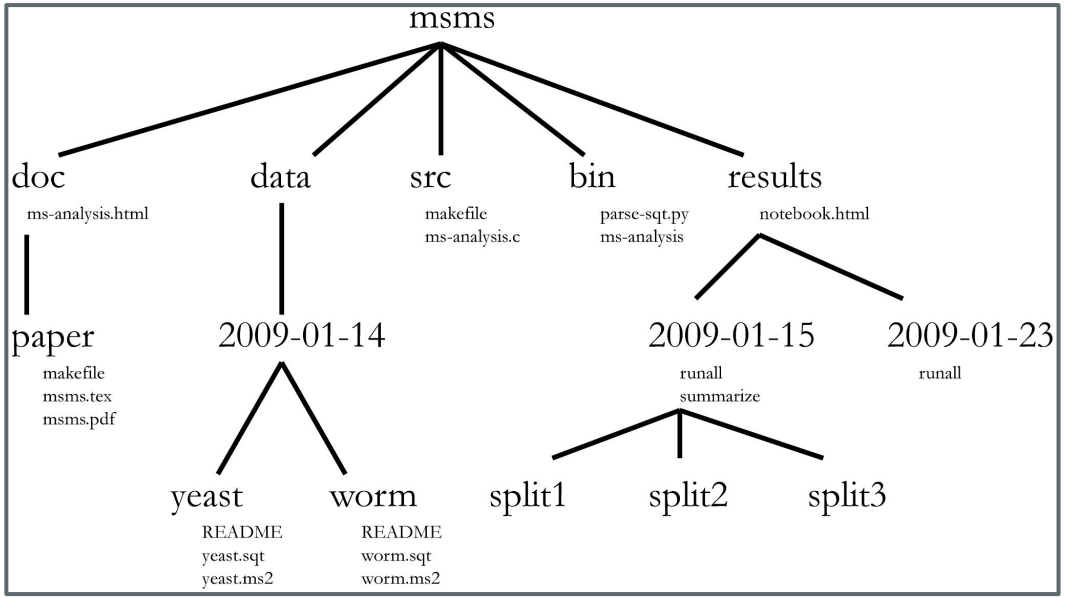
File naming conventions & organization

- Good file names tell you, at a glance, what's in the file
- Organize things based on how you think about them, or mimics how you work.
- In other words: When you look for a file, where do you think it should be?



File naming conventions & organization

- Try to keep scrolling to a minimum
- Try to keep clicks to a minimum
- *Make a system*
- *Document your system*



From A quick guide to organizing computational biology projects, Noble, 2009.
https://doi.org/10.1371/journal.pcbi.1000424

File naming practices

Practice		Example
Limit file names to 32 characters	✔	32CharactersLooksExactlyLikeThis.csv
Stick with alphanumerics (so no special characters or spaces - like . , - $ ())	✔	name_date_location.txt
	☐	name&date@location.txt
	☐	name.date.location.txt
Use versioning	✔	Manuscript_v041.doc
Use leading zeros to help with sequencing	✔	Subj01_analysis; Subj001_processedData
Use descriptive file names (to avoid conflicts if you need to move your files around or share them)	✔	Subj01_raw_20210921.csv
	☐	Mydata.csv

Organizing a lot of data...

Processes & consistency & documentation become even more important.

You may need to lean on renaming and reorganizing programs to get you started, such as:

Bulk Rename Utility (Windows, free)

Renamer (Mac, $$)

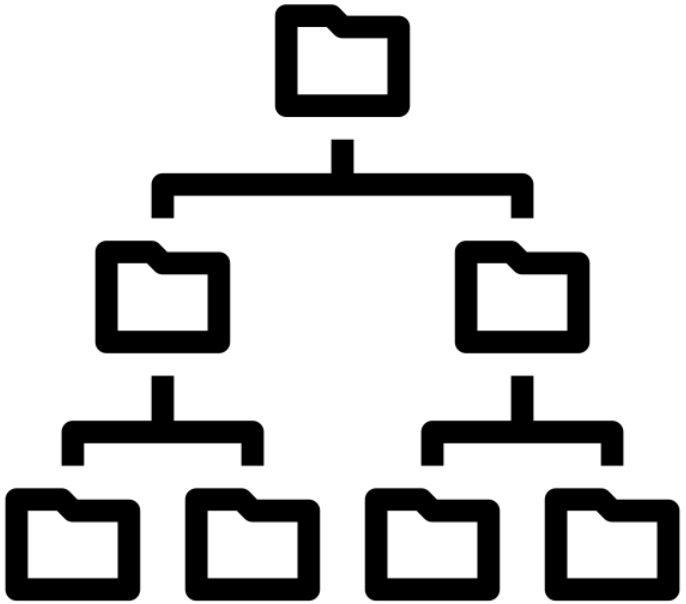
PSRenamer (Linux, Mac, Windows, free)

Additional resources

- MIT's file naming handout: https://www.dropbox.com/s/ttv3boomxlfgiz5/Handout_fileNaming.pdf?dl=0
- MIT's worksheet on naming and organizing your files and folders: https://www.dropbox.com/s/xx26a1onsu1qdpc/Worksheet_fileOrg.docx?dl=0

Take good notes

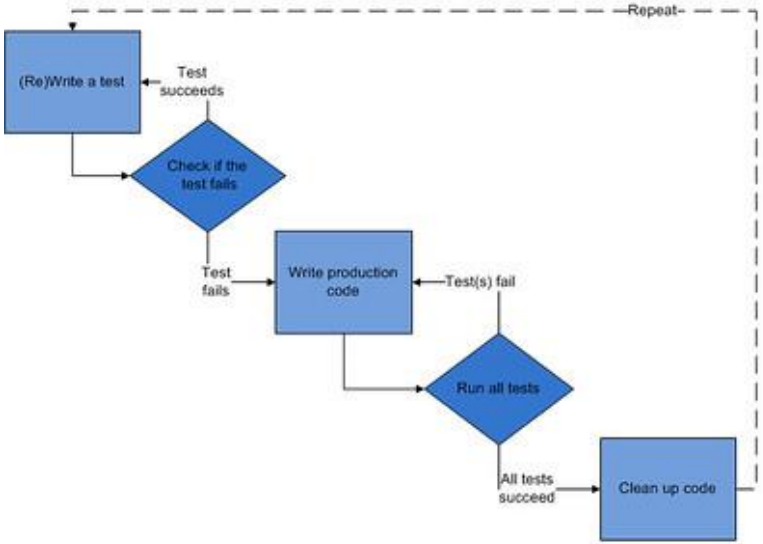
Document your organizational system



Created by monkik
from Noun Project

- What are your file naming conventions?
- What is your folder hierarchy?

Document your workflow



How did you get from raw data to the final product of your research?

http://en.wikipedia.org/wiki/Test-driven_development

Document your data


```
{"results":[{"url":"http://scholarworks.umass.edu/dissertations_2/1948","parent_key":"51
Impact on Learning for Formerly Incarcerated Adolescents in the Age of Zero Tolerance Po
viewcontent.cgi?article=3007&context=dissertations_2","document_type":["openaccess","Ope
scholarworks.umass.edu/dissertations_2/1855","parent_key":"5111417","context_key":"16059
CASE STUDY","publication_date":"2020-02-01T08:00:00Z","download_link":"https://scholarwo
"Campus-Only Access for Five (5) Years"],"author":["Nhu Nguyen"]},{"url":"http://scholar
convergent retrieval learning theory of testing effects","publication_date":"2020-03-24T
context=dissertations_2","document_type":["openaccess","Open Access Dissertation","Open
1990","parent_key":"5111417","context_key":"19165765","title":"FILAMENTS, FIBERS, AND FO
scholarworks.umass.edu/cgi/viewcontent.cgi?article=3126&context=dissertations_2","docume
{"url":"http://scholarworks.umass.edu/dissertations_2/2056","parent_key":"5111417","cont
"publication_date":"2020-09-01T07:00:00Z","download_link":"https://scholarworks.umass.ed
for One (1) Year"],"author":["Rodrigo Mercado Fernandez"]},{"url":"http://scholarworks.u
MASSIVE, EXPENSIVE, OR OTHERWISE INCONVENIENT GRAPHS","publication_date":"2020-12-18T08:
context=dissertations_2","document_type":["openaccess","Open Access Dissertation","Open
"parent_key":"5111417","context_key":"17889913","title":"Lisa Ben and Queer Rhetorical R
scholarworks.umass.edu/cgi/viewcontent.cgi?article=3043&context=dissertations_2","docume
Litterer"]},{"url":"http://scholarworks.umass.edu/dissertations_2/1687","parent_key":"51
High-Intensity Muscle Contractions In Vivo","publication_date":"2019-10-30T15:28:20Z","d
"document_type":["openaccess","Open Access Dissertation","Open Access Dissertations"],"a
"parent_key":"5111417","context_key":"15969084","title":"Optimal Linearization: Prosodic
scholarworks.umass.edu/cgi/viewcontent.cgi?article=2903&context=dissertations_2","docume
"query_meta":{"total_hits":9,"start":0,"limit":100,"field_params":{"include_only":["auth
```


What is that data field? What are the units? What does that acronym mean?

README files: what we need to know to use your data!

- Where to find it
- How to access it
- What can it be used for?
- Known problems, inconsistencies, limitations
- Collection methods, units of measure, variable names
- Fixity checks
- Ethical/privacy restrictions
- Licensing
- Who to cite

README example

https://scholarworks.umass.edu/data/136/

```
This README.txt file was generated on 20210703 by Kelly McKeon
-------------------------------------
GENERAL INFORMATION
-------------------------------------
Title: Freshwater Tidal Wetland Sediment Flux in the Hudson River, NY


##Author Information

        First Author/Corresponding Author Contact Information
                Kelly McKeon
                Woods Hole Oceanographic Institution
                kmckeon@whoi.edu

        Other Author Contact Information
                Jon Woodruff
                University of Massachusetts Amherst
                woodruff@umass.edu

Keywords: tidal wetland, marsh, sediment flux, sediment budgets

Description: This study primarily used a 16-year tidal flux dataset generated by the Hudson River National Estuarine Research Reser
sediment budgets for two freshwater tidal wetlands in the Hudson River. Throughout this dataset, Tivoli North Bay is a marsh and fi
Tivoli South Bay is a mudflat and files associated with this bay will be labelled TVS. The HRNERR dataset is publicly available thr
at cdmo.baruch.sc.edu. To supplement the publicly available water level, turbidity, and suspended sediment concentration (SSC) data
water level loggers at openings to each bay, p
ADCP current measurements are included here, a
of these data with the HRNERR data to estimate

-------------------------------
DATA & FILE OVERVIEW
-------------------------------

## Data Description
All data in this repository are from the Tivol
associated with Tivoli South Bay is labelled T
labelled TVN_N correspond to the northern culv
and TVN means the southern culvert in TVN. Sto
Tivoli South, and is identified in file names

## Data Locations
TVN N: 42.045595, -73.924827
TVN S: 42.036764, -73.925383
TVS N: 42.026742, -73.925970
TVS S: 42.012152, -73.926887
Marsh 1: 42.042490, -73.919530
Marsh 2: 42.045000, -73.921510
M7: 42.023670, -73.921720
M8: 42.017330, -73.923330
```

```
## Folder: Tiltmeter
Lowell Instruments TC-4 Tiltmeters were deployed from July 21, 2020 ñ October 21, 2020, with data downloaded once mid-deployment on August 19, 2020. Current speeds are reported
in cm/s and headings are reported in degrees. Data was collected at one-minute intervals.

## Folder: SSC_LOI
Suspended sediment concentration (SSC) measurements were collected by HRNERR and are publicly available through the HRNERR data management office at  cdmo.baruch.sc.edu. Data in
the SSC_LOI folder contains the clastic fraction of the SSC measurements collected by HRNERR from 2016-2019. Filters were dried, weighted, and burned for four hours at 550C to
combust organics. The final weight divided by the dry sediment weight represents the clastic fraction.

## Folder: Sediment Traps
Sediment traps were installed in two transects across the marsh platform in Tivoli North, with five stations in each transect. At each station, five 50mL, 2.7cm diameter,
centrifuge tubes were pushed into the sediment until level with the marsh platform. Traps at Transect 1 were deployed from June-October, while traps at Transect 2 were deployed
from August-October. Sediments collected in the traps underwent loss-on-ignition following the same method described above for water column filters. Clastic sediment mass was
divided by the surface area of the trap and the length of time deployed to calculate the deposition rate over each trapís deployment period, and we report the average of those
rates in g/cm2/yr.

## Folder: Hobo
Onset HOBO water level loggers were deployed from July 21, 2020 ñ October 21, 2020, with data downloaded once mid-deployment on August 19, 2020. Both culverts in Tivoli North,
the northermost and southernmost culverts in Tivoli South, and two locations on the marsh platform were monitored. Files labelled ìMarshî correspond to loggers deployed on the
marsh platform. The HOBO used to measure barometric pressure was deployed at the southern culvert in Tivoli South and is labelled ìBaroî in the files. Data was collected at 15-
minute intervals. Pressures are reported in kPa and temperature is reported in Celsius.

## Folder: Gamma
Sediment cores were collected in two locations in Tivoli South and Cs-137 was measured via gamma spectroscopy. The upper meter of each core was sampled every 10 cm for ~2 cm3 of
sediment. Samples were dried, crushed, and counted for at least 48 hours on a Canberra GL2020R Low Energy Germanium Detector. Units are in Bq/g.

## Folder: ADCP
ADCP surveys were conducted at all culverts on October 19, 2020 over the course of one tidal cycle. The ADCP was passed between two to five times on the bay side of each culvert
over time frames ranging from twenty minutes to one hour. VMT files from each pass are in folders corresponding to their culverts.
```

README best practices

- Create one README file for each data file/dataset
- Name the README so that it's easily associated with the file(s) it describes
- Write your README document as a plain text file
- Format multiple README files identically
- Follow the conventions for your discipline

Resources:

- https://scholarworks.umass.edu/data/guidelines.html
- https://guides.library.umass.edu/data/share/prepare

Security

Security: Why you should care

- Personally identifiable information
- Protected populations (human/non)
- Controlled unclassified information
- Data use agreements
- Human subjects
- Proprietary data
- Patents

Security options

Passwords & Password managers (KeePass, SplashID, 1Password, Keychain (MacOSX))

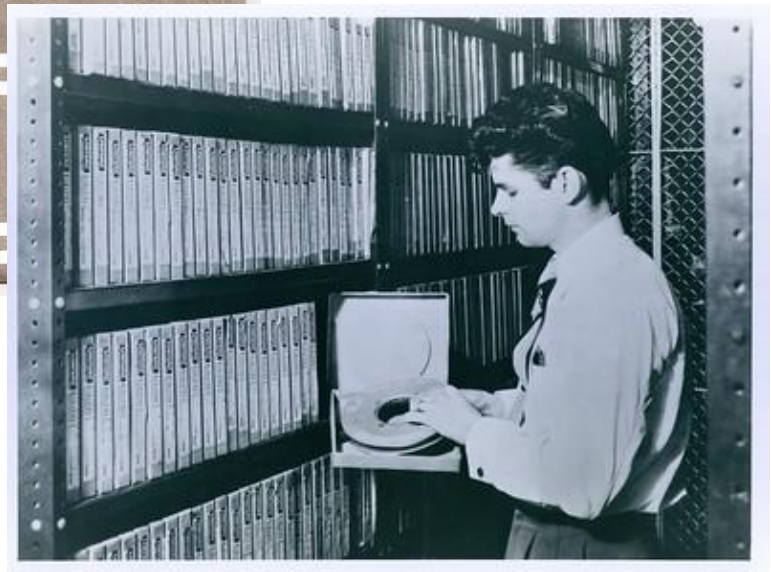
UMass Data Protection Action Plan: https://www.umass.edu/it/support/security/data-protection-action-plan

Encryption: https://www.umass.edu/it/security/data-encryption-umass-amherst

Future proof

Lifespan of Storage Media

Media	Estimated Lifespan
Magnetic data (tapes)	Up to 10 years
Nintendo cartridge	10-20 years
Floppy disk	10-20 years
CDs and DVDs	5-10 unrecorded, 2-5 recorded
Blu-Ray	Not certain, probably over 2-5 recorded
M-Disc	1,000 years (theoretically)
Hard disk	3-5 years
Flash storage	5-10 years or more (depends on write cycles)



https://blog.storagecraft.com/data-storage-lifespan/

File formats

Data type	Preferred file format examples
Containers	TAR, GZIP, ZIP
Databases	XML, CSV
Geospatial	SHP, DBF, GeoTIFF, NetCDF
Moving images	MOV, MPEG, AVI, MXF
Sounds	WAVE, AIFF, MP3, MXF
Statistics	ASCII, DTA, POR, SAS, SAV
Still images	TIFF, JPEG2000, PDF, PNG, GIF, BMP
Tabular data	CSV
Text	XML, PDF/A, HTML, ASCII, UTF-8
Web archive	WARC

File formats

Proprietary Format	Alternative/Preferred Format
Excel (.xls, .xlsx)	Comma Separated Values (.csv or .tsv) ASCII
Word (.doc, .docx)	Plain text (.txt) PDF/A (if formatting is needed)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Photoshop (.psd)	TIFF (.tif, .tiff)
Quicktime (.mov)	MPEG-4 (.mp4)

Converting file formats? Be aware of information loss!

To mitigate the risk of lost information when converting:

- Note the conversion steps you take
- If possible, keep the original file as well as the converted ones

Credits:

- Quick & dirty data management: the 5 things you need to be doing now! by Data Management Services.Copyright © 2020-04-16 MASSACHUSETTS INSTITUTE OF TECHNOLOGY is licensed under a Creative Commons Attribution 4.0International License except where otherwise noted. [https://creativecommons.org/licenses/by/4.0/]. Access at https://www.dropbox.com/s/s34cwxoamzq30im/QuickDirtyDataMgmt_Slides_MIT.pdf?dl=0

Thanks & let's chat!